THYROID DISEASE CLASSIFICATION USING MACHINE LEARNING

**TEAM LEADER MENTOR**

Vanitha S Dr.E.Elakkiya M.sc., M.phil, ph.D

**MEMBERS**

Thilakavathi M

Vithya R

Ramya S

INTRODUCTION

One of the most vital organs in the human body is the thyroid gland, a vascular gland. Two hormones secreted by this gland aid in regulating the body's metabolism. Hyperthyroidism and hypothyroidism are the two different forms of thyroid diseases. They cause the body to release particular hormones, which causes an imbalance in the metabolism. To diagnose this condition, a blood test for the thyroid is used, although the results are frequently distorted and noisy. To make the data simple enough for analytics to highlight the likelihood of patients contracting this condition, data purification techniques were applied. In order to forecast diseases, machine learning is crucial. SVM - Support Vector Machine techniques for machine learning. To determine the patient's likelihood of developing thyroid disease, the Random Forest Classifier, XGB Classifier, and ANN - Artificial Neural Networks - are all used. The purpose of the online application is to gather user information for disease classification.

OVERVIEW

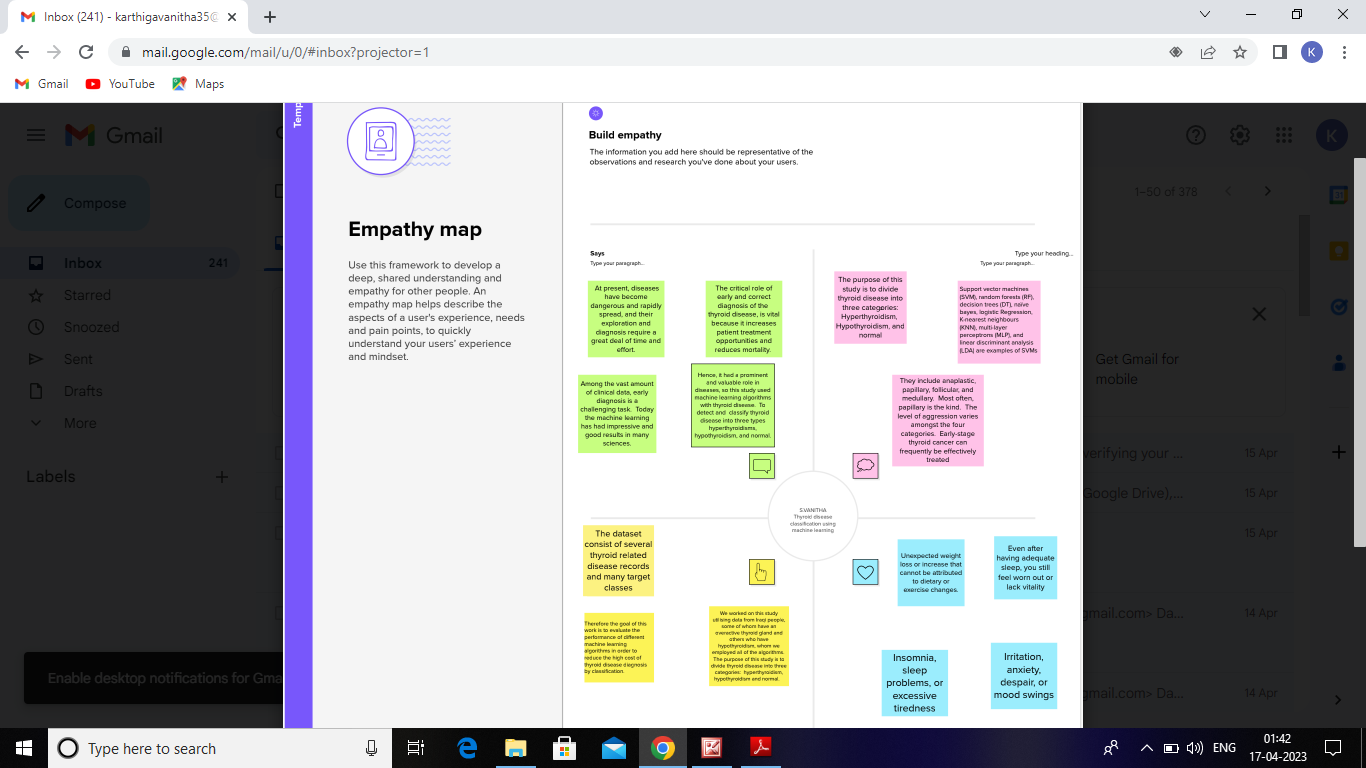
A better knowledge of the cell of origin, pathologic characteristics (cytopathology and histology), molecular categorization, and biological behavior is now possible thanks to the new classification, which has separated thyroid tumours into a number of new groups. The majority of thyroid neoplasms are formed from follicular cells. They are separated into benign, low-risk, and malignant neoplasms under this new classification. In addition to follicular adenoma, benign tumours can also include adenoma types that are significant for diagnosis and treatment, such as those are frequently hyper functional and oncocytic adenoma.

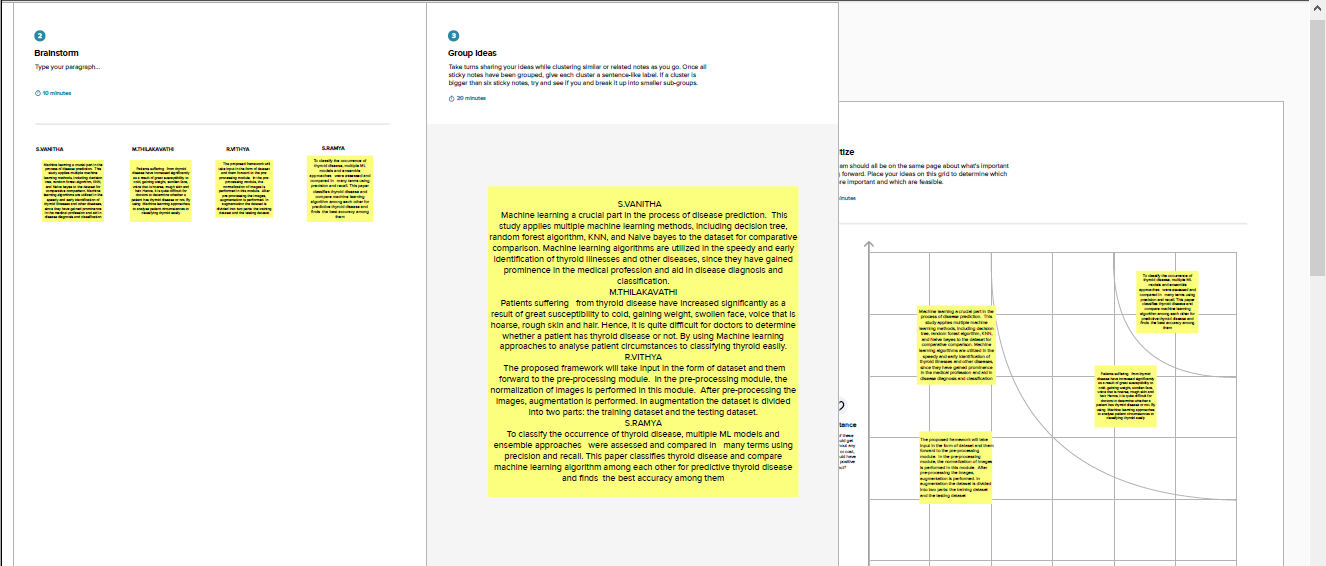
PURPOSE

The application of machine learning techniques may help to enhance the classification of thyroid illnesses since the initial diagnosis of thyroid tumors based on histological patterns can occasionally be confusing. Additionally, the classification problem is complicated further by the dearth of highly sampled datasets.

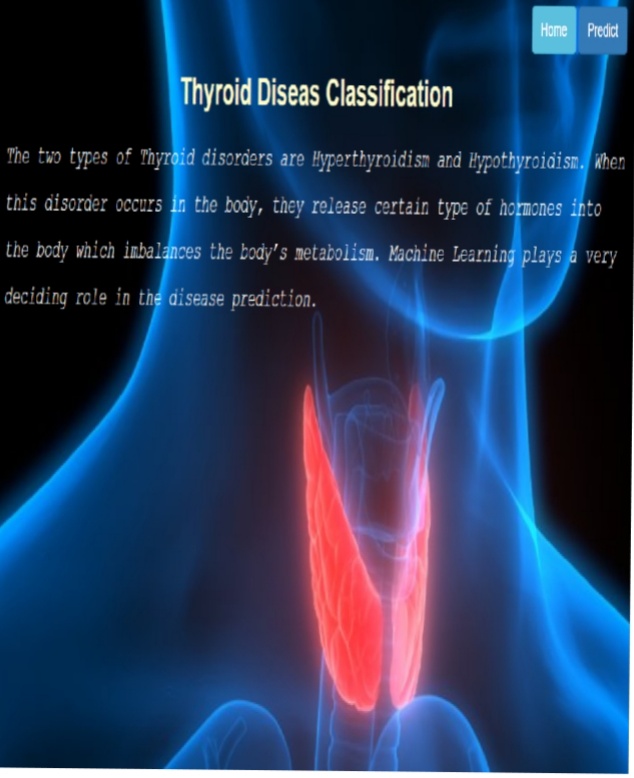
2. PROBLEM DEFINITION & DESIGN THINKING

EMPATHY MAP

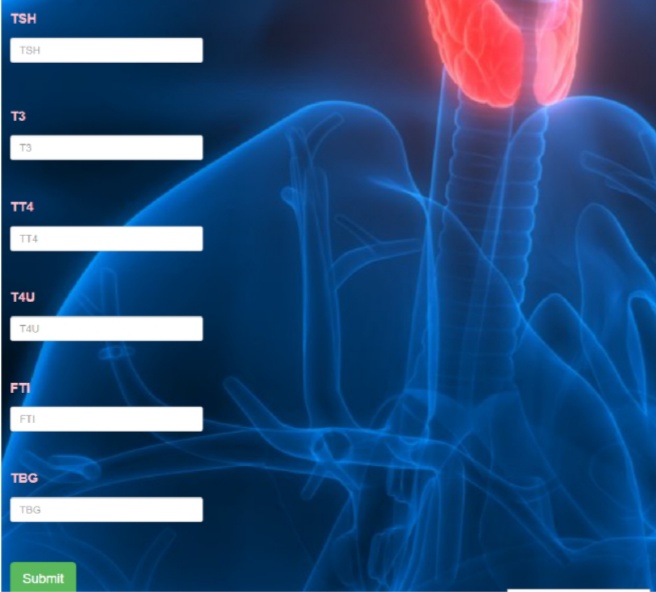


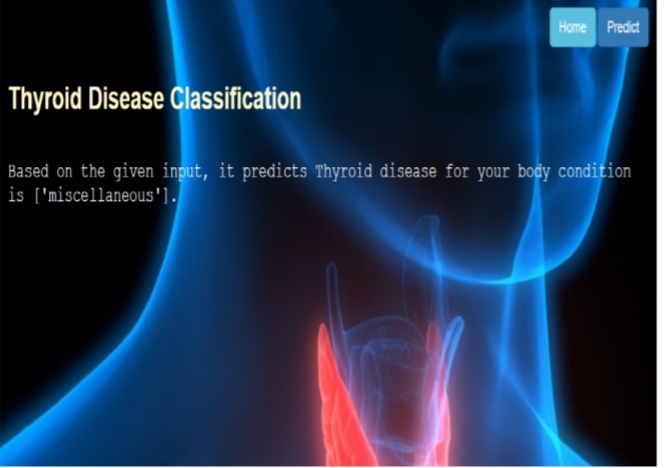


RESULT









ADVANTAGES AND DISADVANTAGES

|  |  |
| --- | --- |
| **ADVANTAGES** | **DISADVANTAGES** |
| Standaraized method  Easy learning curve  Reduced operating time | Poor cosmetic result  Hyper- or paresthesia in the neck |
| Better cosmetic result  Less postoperative pain | Difficult learning curve  Limits in nodules size |
| Scarless technique | Greater postoperative pain  Difficult learning curve  Longer operating time  Complications from  CO2 insufflation  Higher cost |
| Better cosmetic result  Three-dimensional working field  Magnified view  Eliminate tremor | Higher cost  Greater postoperative pain  Longer operating time  Very difficult learning curve |

**Conclusion**

One of the disorders that affects the global population and is becoming more prevalent is thyroid disease. Our study examines the classification of thyroid disease between hyperthyroidism and hypothyroidism in light of medical reports that demonstrate substantial abnormalities in thyroid diseases. Utilizing algorithms, this illness was categorized. Using a variety of methods, machine learning produced positive outcomes. This study shows how data mining and machine learning approaches can help the medical community and healthcare system. This study will assist the medical professionals in using this as an additional system in accordance with standard protocol. We assessed the dataset using recall and precision.

**FUTURE SCOPE**

Thyroid Detection using Machine Learning is a smart and precise way to predict thyroid disease. The first work is

collecting the data from the UCI repository, then analyzing it with exploratory analysis where I found insights from the

data, then the data was cleaned and transformed for prediction. NB, KNN, LightGBM, DT, XGBoost, AdaBoost, and

CatBoost have been implemented and precision, recall, accuracy, F1 score were used to evaluate the implemented

models’ performance. XGBoost classifier did pretty well achieving the highest accuracy. Other evaluation metrics

also support the performance of this algorithm. I thereby recommend the XGBoost classifier for the predictive model.

The running times of the algorithms are compared. CatBoost took the highest amount of time. LightGBM took less time than XGBoost. The optimal model that should be used for this dataset is LightGBM for fast results and XGBoost for a igher accurate model. To predict thyroid disorder, all classifiers produce good results except NB. In the future, these algorithms can be implemented for the prediction of thyroid disease with more real data related to thyroid and with multiple classes

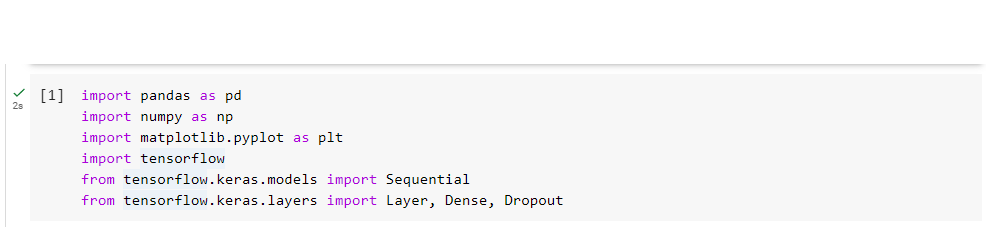
Data Collection & Preparation

ML significantly relies on data. It is the element that enables algorithm training the most. Therefore, you are able to obtain the needed dataset from this section.

Collect the dataset

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

Importing the libraries



Pandas

Pandas is an open-source library designed primarily for dealing quickly and logically with relational or labelled data. It offers a range of data structures and functions for working with relational or labelled data. It offers a range of data structures and functions for working withtime series and numerical data. Pandas is quick and offers its customers high performance & productivity.

NumPy

Numerical Python is referred to as NumPy. The Python library NumPy is used to manipulate collections. Additionally, it has matrices, fourier transform, and tools for working in the area of linear algebra.

Matplotlib

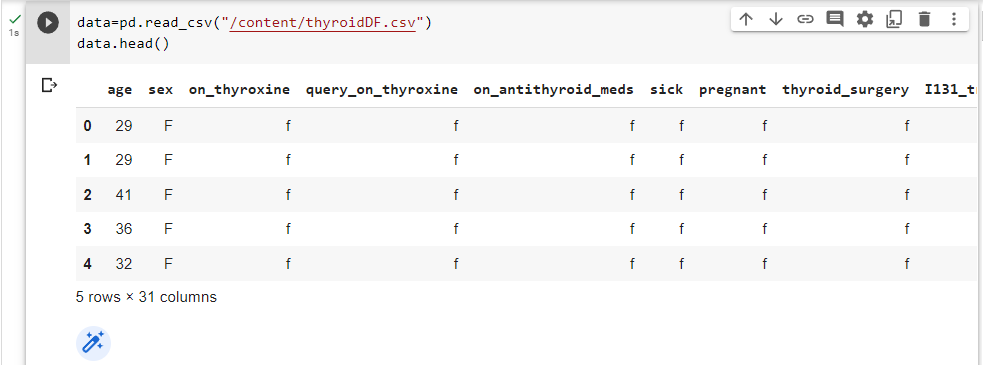
Matplotlib is a fantastic Python visualisation tool. A multi-platform data visualisation library called Matplotlib was created to operate with the larger SciPystack and is based on NumPy arrays. There are numerous graphs in Matplotlib, including line, bar, scatter, histogram,etc.

TensorFlow

A free and open-source software framework for artificial intelligence and machine learning is called TensorFlow. It can be used for a variety of activities, but focuses particularly on deep neural network training and inference.

Read the Dataset

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas. In pandas, we have a function called read\_csv() to read the dataset. As a parameter , we have to give the directory of the csv file.



Data Pre-processing

Now that we know how the data works, let's pre-process the information that was gathered.We need to properly clean the dataset in order to get excellent results because the downloaded data set may contain too much randomness to be used for machine learning model training. The stages in this activity are as follows.

• Handling Categorical Values

• Checking Correlation

• Converting Data Type

• Splitting the dataset into training and test group

• Handling Missing Values

• Descriptive Analysis

• Splitting the dataset as x and y

• Handling Missing Values

• Handling Missing Values

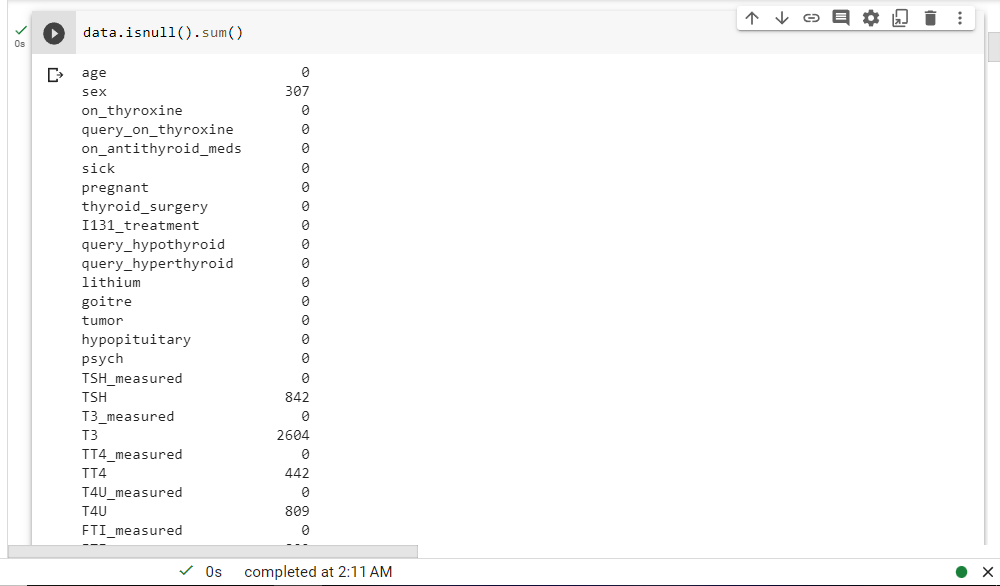
• Handling Imbalanced Data

• Applying StandardScaler

The basic procedures for pre-processing data before using it for machine learning are as follows. You might or might not need to follow each stage, depending on the state of your dataset.

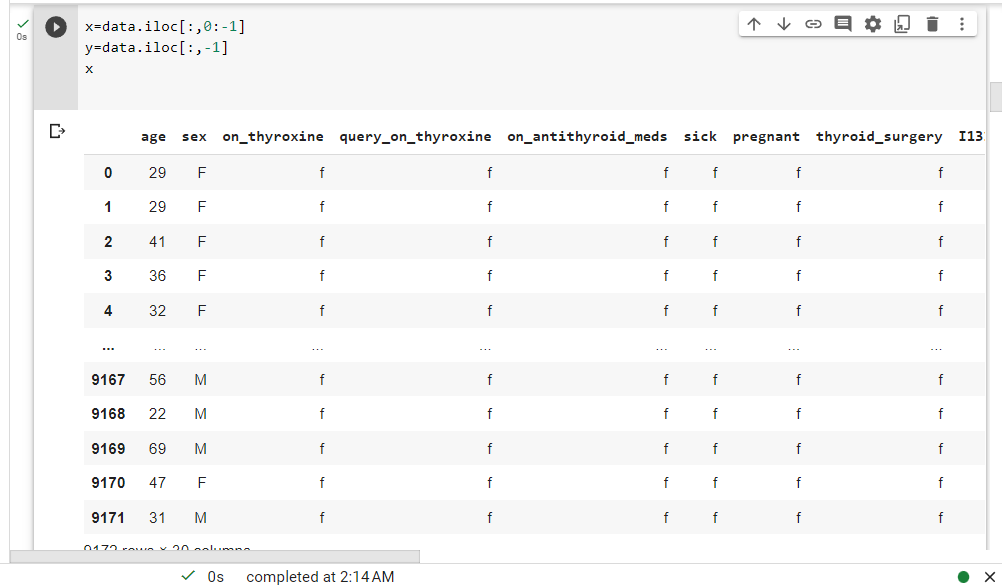
**Checking for null values**

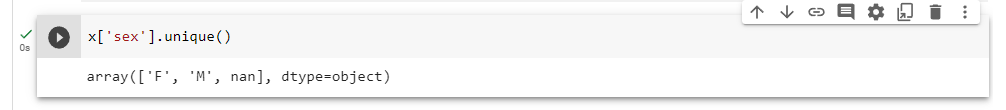
For checking the null values, data.isnull() function is used. To sum those null values we use the sum() function to it. From the below image we found that there are no null values present in our dataset. So we can skip handling the missing values step

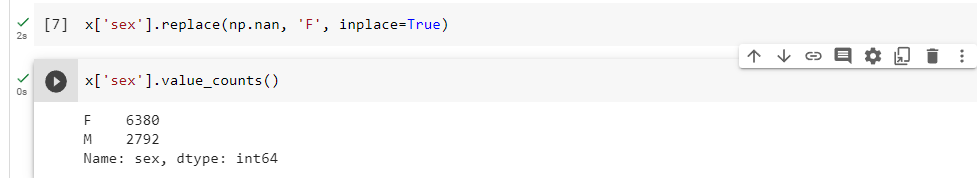


Splitting the data x and y

Splitting the data x and y



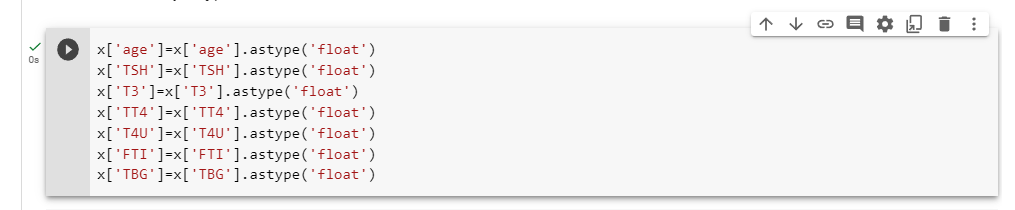




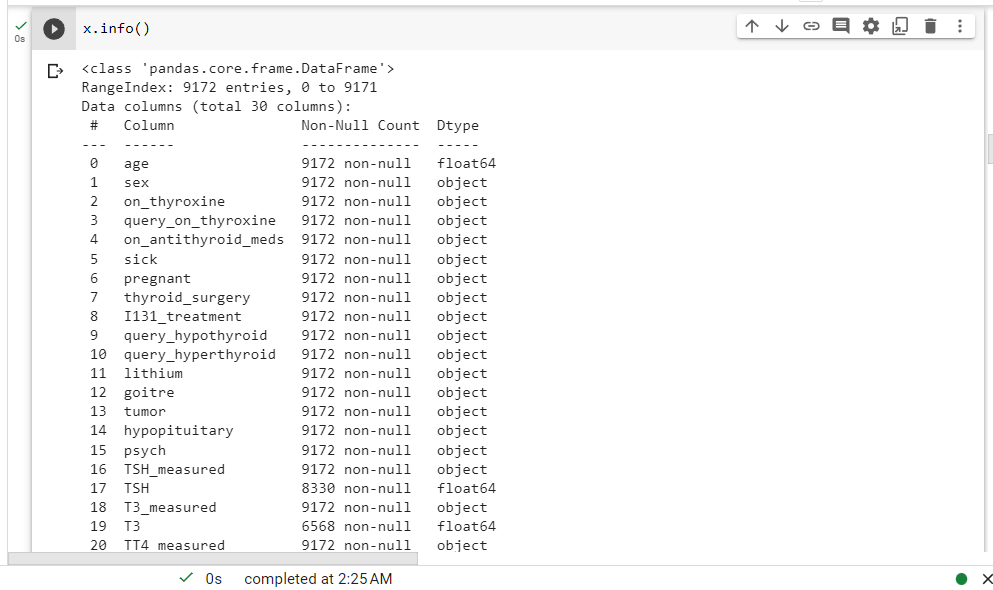
Converting the data type

Converting an object data type to a float data type. So that the output would be accurate and verifying data information

Here, we convert the object values of "TSH," "T3," "TT4", "T4U," "FTI," and "TBG" to float values.



Then we can check the data type information about the dataset by code of x.info()



Handling categorical values

As we can see, our dataset contains categorical data, hence we must encode the categorical data as binary or integer data.

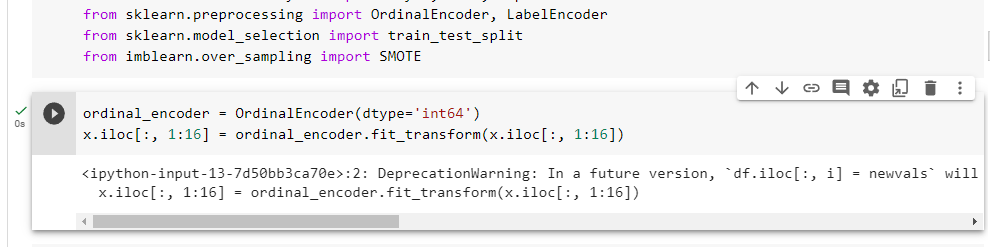
Encoding is used to transform categorical characteristics into numerical features.

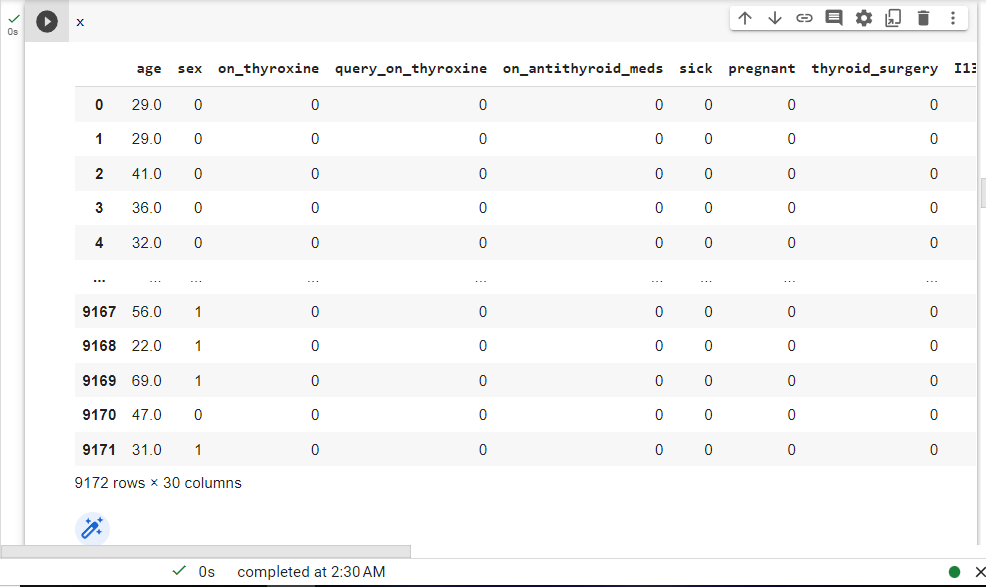
Techniques

Although there are several approaches, we are employing Ordinal Encoding and Label Encoding in our project.

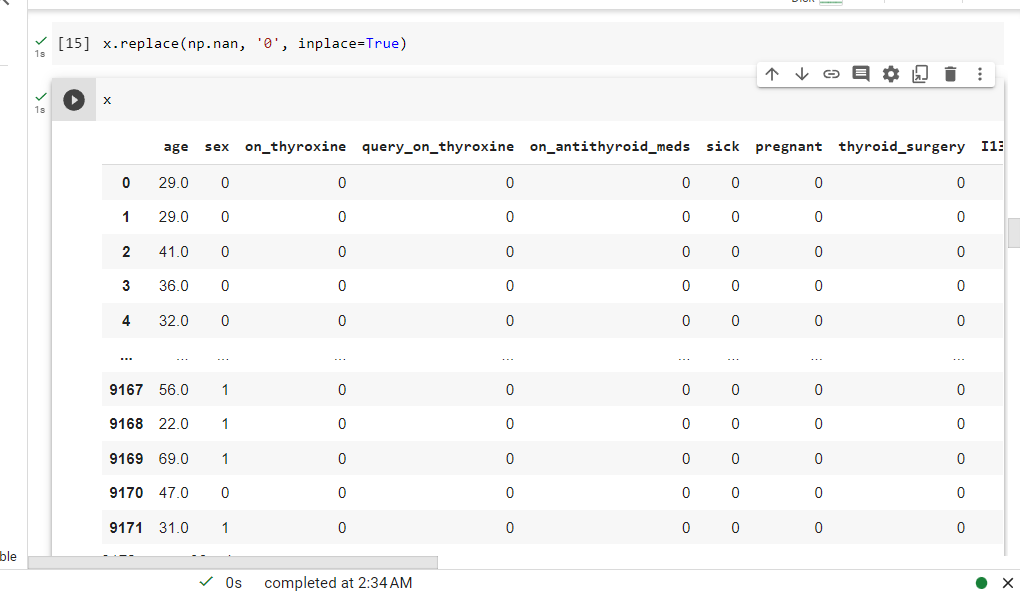
• X and Y values are categorical characteristics in our project.

• X values in this case using Ordinal Encoding.

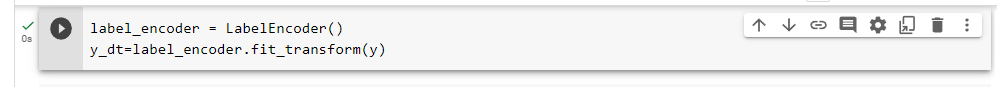


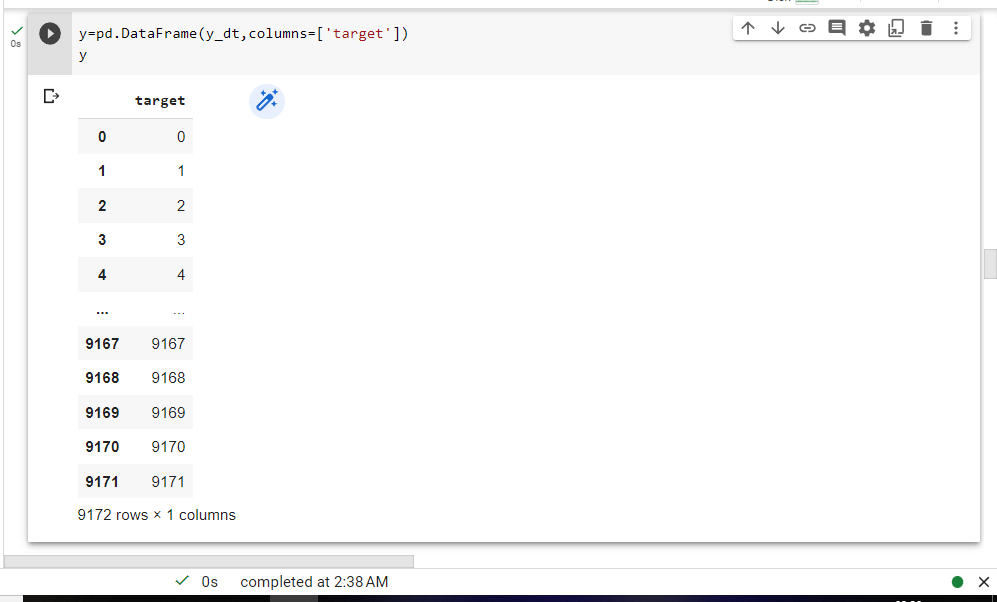


Replacing the non values with zero values



Now applying Label Encoding on y value





Splitting data into train and test

Now let’s split the Dataset into train and test sets

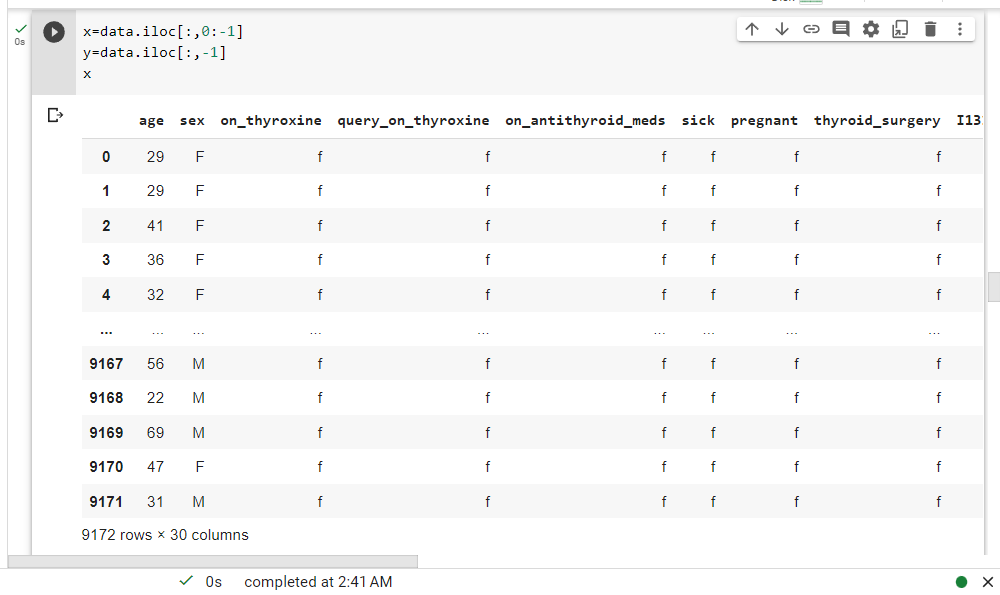
Changes: first split the dataset into x and y and then split the data set

Here x and y variables are created. On x variable, data is passed with dropping the

target variable. And my target variable is passed. For splitting training and testing

data we are using the train\_test\_split() function from sklearn. As parameters, we

are passing x, y, test\_size, random\_state.





Handling Imbalanced Data

Applying Standard Scalar

Scaling the features smooth’s the gradient descent process and speeds up the algorithms' arrival at the cost function minima.

• If scaling characteristics are absent, the algorithm may favor the feature.

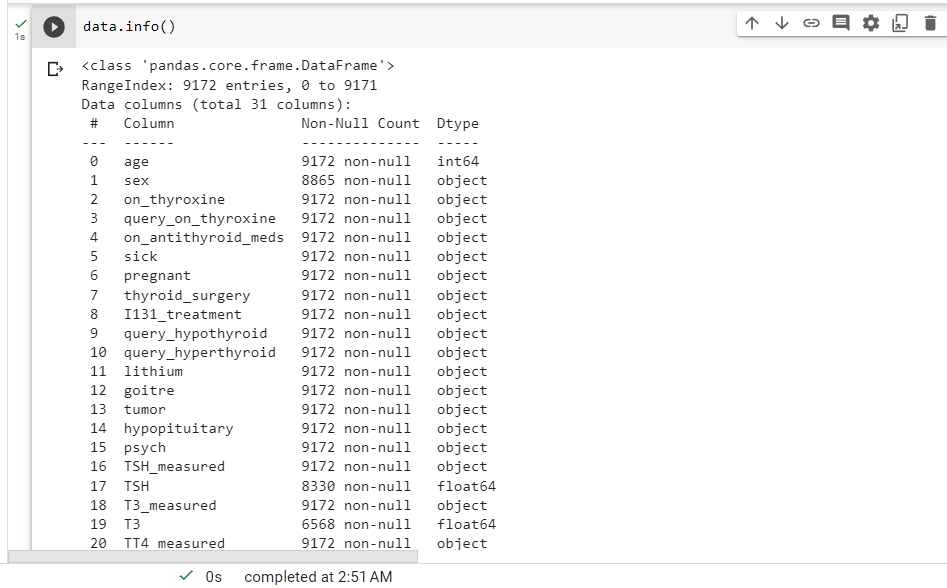
It has greater magnitude values. Every feature is brought into the same range, and the model makes intelligent use of each one.

• In this case, we are converting the array-formatted data into a data frame (tab)

Exploratory Data Analysis

Descriptive analysis

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas have a worthy function called describe. With this described function we can find Checking info about data by using data info().mean, std, min, max and percentile values of continuous features.

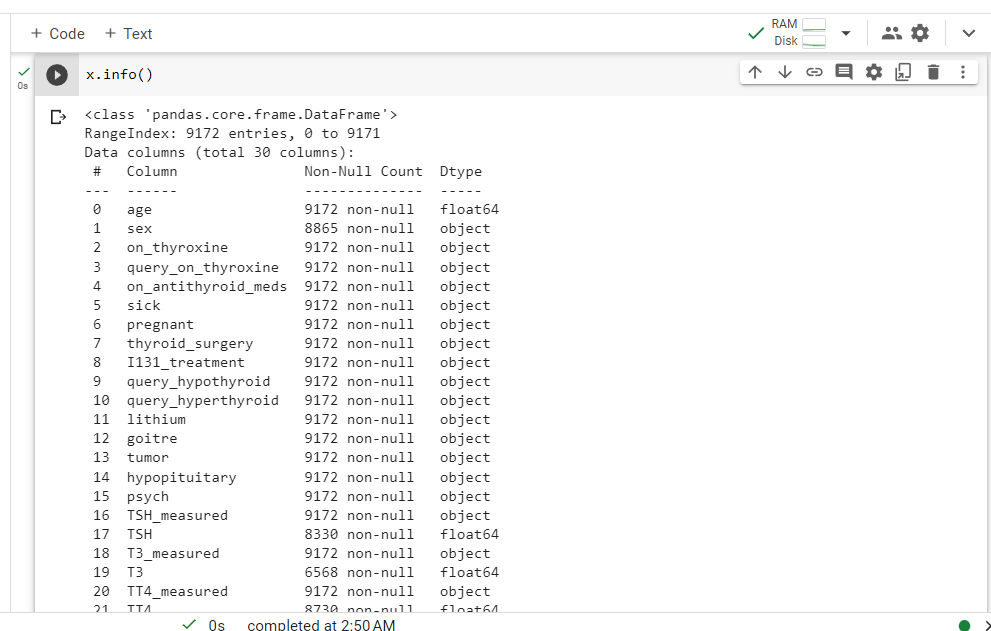


Exploratory Data Analysis

Descriptive Analysis

With the help of statistics, descriptive analysis examines the fundamental characteristics of data. In this case, the describe function in pandas is excellent. This described function allows us to find Continuous feature values include mean, standard deviation, min, max, and percentile.

Using data info () to verify data information



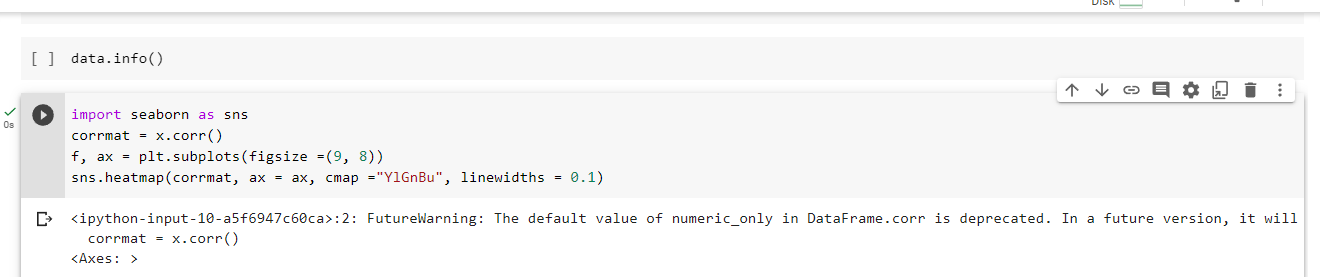
Visual Analysis

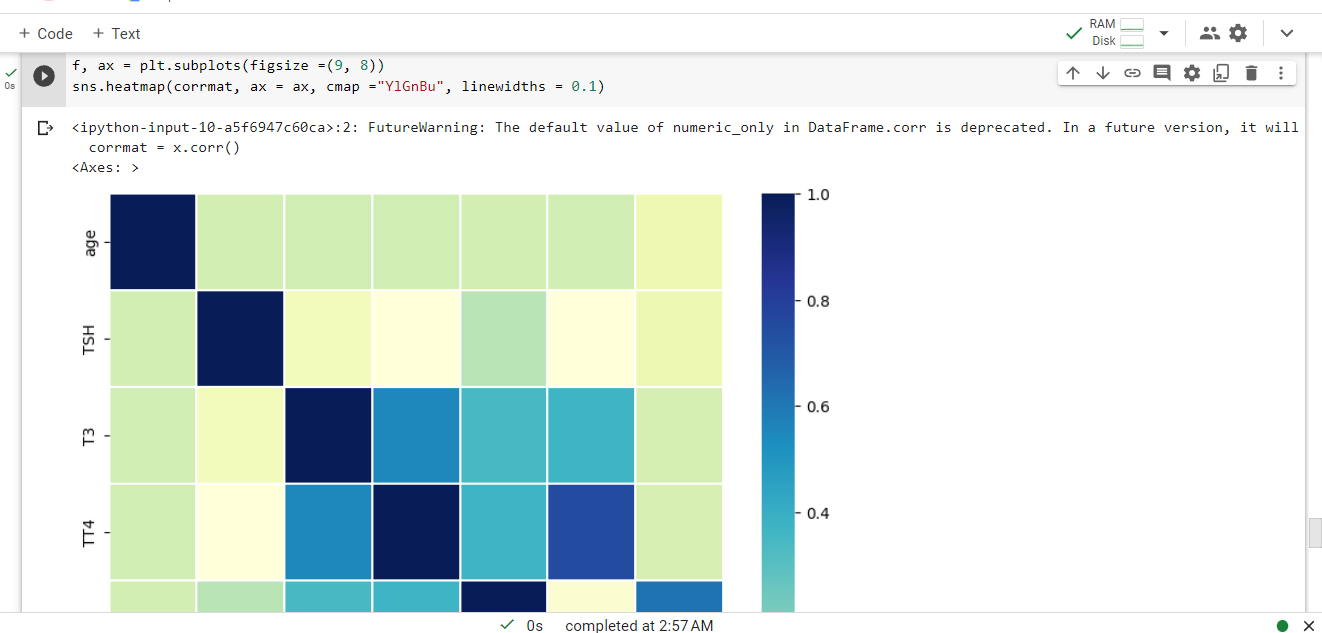
Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify Patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

Checking Correlation.

I'm using Heat Map to discover the association in this case. It uses color variations to show the data as 2-D colored maps. It will be plotted on both axes and uses colors rather than numbers to describe the connected variables.

There is no correlation between the columns in this case





Model Building

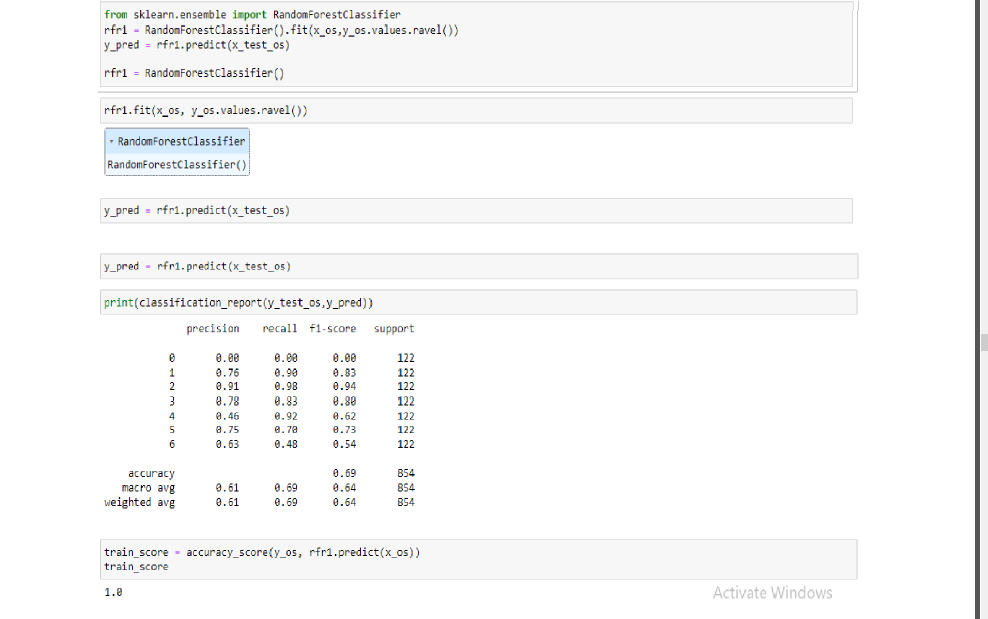
Training the Model in Multiple Algorithms

Now that our data has been cleansed, it's time to construct the model. Our data can be used to train a variety of algorithms. We are using four classification methods for this project. Depending on how well it performs, the best model is saved.

Random Forest Classifier Model

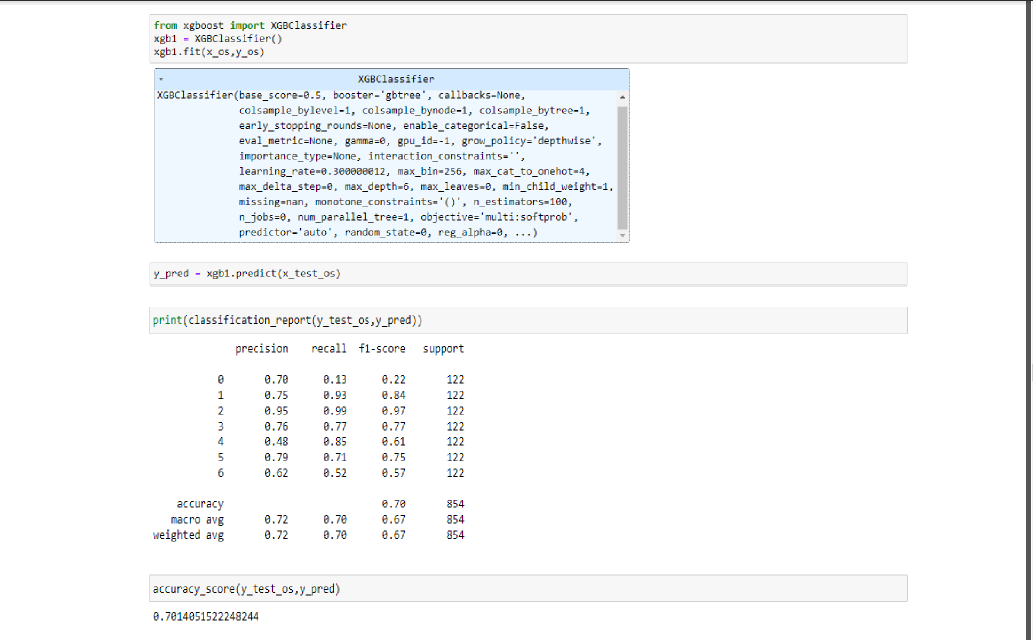
The parameters of a function called Random Forest Classifier Model are train and test data. The function initializes the Random Forest Classifier algorithm and provides the model with training data fit() method.

The predict() function is used to predict test data, which is then saved in a new variable. The accuracy score and classification report are completed for the model's evaluation.



XGB Classifier Model

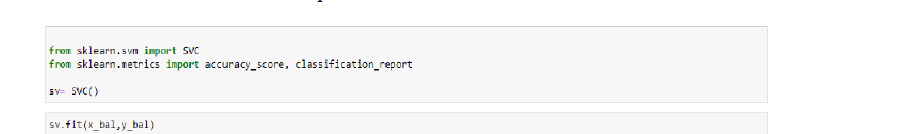
Train and test data are supplied as parameters to a function called XGB Classifier model. The XGB Classifier algorithm is initialized inside the function, and training data is supplied to the model using the .fit () function. The test predict () function is used to forecast data, which is then saved in a new variable. The model is assessed using the accuracy score and classification report.

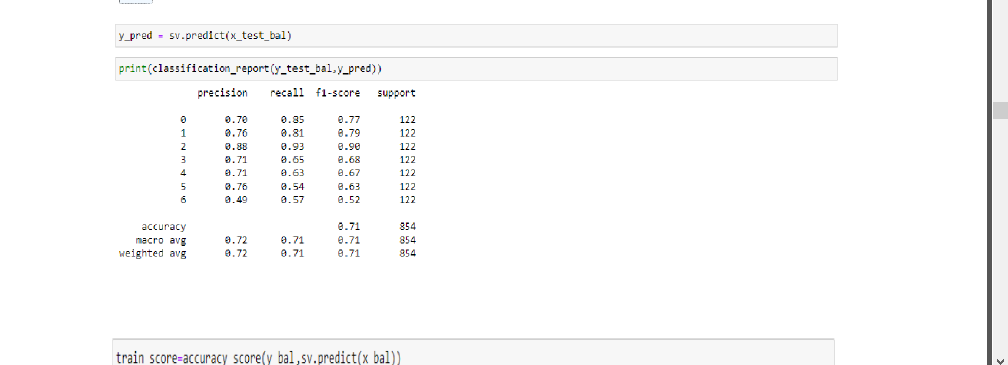


SVC Model

Train and test data are supplied as parameters to a function called SVC model. The SVC algorithm is set up inside the function, and the model is given training data using the .fit() method. With the .predict () function, test data is predicted.

A new variable and saved in a new function. The accuracy score and classification report are completed in order to evaluate the model.





ANN Model

Multi-layer, fully-connected neural nets are what artificial neural networks (ANN) are made of.

An input layer, a number of hidden layers , and an output layer make them up. Each node in one layer is linked to each node in the layer above it. Making the deepen the network by adding more covert layers

